Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments

Stefan E Seemann Jan Gorodkin Rolf Backofen

Supporting Information

SCFG for the Pfold model

Let A be an alignment, and let $\vec{A}^1, \ldots, \vec{A}^m$ be the tuple of columns of A, where m is the length of the alignment A, and $\vec{A}^i$ is the ith column of A. The probability distribution on structure $\Pr[\sigma \mid A, T, M]$, given the data (i.e., the multiple alignment A of the sequences $s_1 \ldots s_n$) and the background information (i.e., the secondary structure background model M and the tree T), is calculated with a combined SCFG by the Pfold model. Let $\tau_M(\sigma)$ be the associated parse tree that produces the structure σ using the grammar M. For each node n in $\tau_M(\sigma)$, let label(n) be the associated terminal or non-terminal symbol, rule(n) the associated grammar rule producing this node, and $\mathrm{pos}(n) = (i, j)$ the pair of start and end position of the produced sequence covered by the node (i.e., the leafs below n is the sequence $s_i \ldots s_j$). Denote with $A_{(i,j)}$ the corresponding sub-alignment. Furthermore, let $n_1 \ldots n_k$ be the children of n. Then we recursively define

$$\Pr_{\tau_M(\sigma)}(n, A_{\mathrm{pos}(n)}) =$$

$$\Pr[\mathrm{rule}(n)|M] \times \prod_{\ell=1}^{k} \Pr_{\tau_M(\sigma)}(n_\ell, A_{\mathrm{pos}(n_\ell)}) \times \begin{cases} \Pr_{\mathrm{bp}}[\vec{A}^i \vec{A}^j|T] & \text{if rule}(n) = F \to dFd \\ & \text{or rule}(n) = L \to dFd \\ \Pr_{\mathrm{sg}}[\vec{A}^i|T] & \text{if rule}(n) = L \to s \\ 1 & \text{else} \end{cases}$$

where $\Pr_{\mathrm{bp}}[\vec{A}^i \vec{A}^j|T]$ and $\Pr_{\mathrm{sg}}[\vec{A}^i|T]$ are calculated in Pfold using Felsensteins's dynamic programming for phylogenetic trees. In principle, it is just the recursive definition of the probability of a parse tree given a grammar, extended by position specific probabilities for producing the terminals. For nodes n that are leaves one defines $\Pr_{\tau_M(\sigma)}(n, A) = 1$. Finally, we define

$$\Pr[A|T, \sigma, M] P[\sigma|T, M] = \Pr_{\tau_M(\sigma)}(\mathrm{r}(\sigma), A),$$

where $\mathrm{r}(\sigma)$ is the root node of $\tau_M(\sigma)$. Since we are not using the parse tree $\tau_M(\sigma)$ explicitely in the main text, we will write $\Pr(\mathrm{r}(\sigma), A)$ as short for $\Pr_{\tau_M(\sigma)}(\mathrm{r}(\sigma), A)$.

As shown in equation 1, $\Pr_{\tau_M(\sigma)}(\mathrm{r}(\sigma), A)$ and $\Pr[\sigma \mid A, T, M]$ differ only by a factor $\Pr[A \mid T, M]$ which is independent from the structure σ. Hence, we have

$$\operatorname*{argmax}_\sigma \Pr[\sigma \mid A, T, M] = \operatorname*{argmax}_\sigma \Pr_{\tau_M(\sigma)}(\mathrm{r}(\sigma), A)$$

Nussinov algorithm

PETfold uses a Nussinov style algorithm to calculate the consensus structure of an alignment with maximal expected overlap. The Nussinov algorithm uses dynamic programming to find the structure with the highest score. Let $F(i, j)$ denote the maximal score of an RNA structure for the sequence $s_i \ldots s_j$. Thus, we have

$$F(i,j) = max \begin{cases} F(i+1, j) + s(x_i) \\ F(i, j-1) + s(x_j) \\ F(i+1, j-1) + s(x_i, x_j) \\ max_{i \le k < j} \{F(i,k) + F(k+1, j)\} \end{cases}$$

where $s(x_i)$ (and $s(x_j)$) is the score for a single-stranded position x_i and $s(x_i, x_j)$ is the score for paired bases x_i and x_j. In PETfold the single-stranded score of position x_i consists of the evolutionary reliability ${\mathcal{R}^{\text{sg}}}_{A,T,M}(i)$ and the thermodynamic probability $\frac{1}{n}\sum_{\text{u}} q^u_{f_A^{-1}(i)}$ over all sequences s_u $(1 \leq u \leq n)$ in the alignment, and the base pair score of the positions x_i and x_j consists of the evolutionary reliability $\mathcal{R}_{A,T,M}(i,j)$ and the thermodynamic probability $\frac{1}{n}\sum_{\text{u}} p^u_{f_A^{-1}(i,j)}$. The optimal structure σ can be reproduced by backtracking from $F(1,L)$ when L is the sequence length. In PETfold, we define ex-over$(\sigma) = F(1,L)$.

Calculation of structural entities with high reliability

We present a statistical method to estimate reliability thresholds for conserved functional regions. Single stranded positions and base pair positions are collected that have a high evolutionary reliability. We write down only the base pair part. Single-stranded positions are treated analogously. For this purpose, do the following

1. Generate shuffled alignment A^{shuffle} by shuffling the alignment columns. Then, we generate again the most likely structure under the shuffled alignment, i.e., we generate

$$\sigma^{\text{MAP}}_{A^{\text{shuffle}},T,M}$$

Then, we collect all the reliability scores for base pairs that are contained in this structure, and iterate this several times:

$$\mathcal{B} = \biguplus_{A^{\text{shuffle}}} \left\{ \mathcal{R}_{A^{\text{shuffle}},T,M}(i,j) \mid (i,j) \in \sigma^{\text{MAP}}_{A^{\text{shuffle}},T,M} \right\}$$

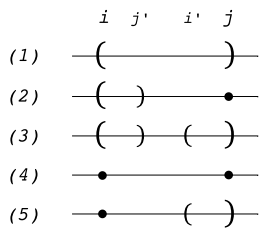
Finally, we order them in size $p_1 > p_2 > \cdots > p_{|\mathcal{B}|}$ and select a significance level θ (e.g., $\theta = 0.01$). Then the probability $p_{\lceil \theta |B| \rceil}$ is the base pair probability $p^{\text{threshold}}$ such that any base pair $(i,j) \in \sigma^{\text{MAP}}_{A^{\text{shuffle}},T,M}$

$$\Pr[\mathcal{R}_{A^{\text{shuffle}},T,M}(i,j) > p^{\text{threshold}}] \leq \theta$$

We applied the previously described stepwise approach on our data set consisting of 46 RNA families. We shuffled for each family 1000 times with a conservative method which mononucleotidely shuffles only columns with the same pattern of gaps and conservation. Then we averaged over the significance values of all families. Using a significance level $\theta = 0.01$, we got a threshold for high reliable base pairs of $p^{\text{threshold}}_{bp} = 0.985$ and single-stranded positions $p^{\text{threshold}}_{ss} = 0.987$, as well as $p^{\text{threshold}}_{bp} = 0.914$ and $p^{\text{threshold}}_{ss} = 0.959$ using $\theta = 0.1$. However, the parameter tuning has indicated that the performance of reliability thresholds depend on another parameter (the weighting factor for single-stranded positions α) which has high impact in the RNA structure prediction of PETfold, and that slightly different reliability thresholds perform better for the data set.

R_5 correlation coefficient

Given two structures in bracket notation, a more stringent secondary structure evaluation can be carried out by considering all pairs of positions, and evaluate the agreement in their structural notation (i.e., dots, opening and closing brackets) in both structures. For each pair of positions (i,j), there are five possible cases. The two positions can be unpaired (4) or paired with each other (1). Furthermore, only the left (2) (resp. right (5)) position can have an opening (resp. closing) bracket. Finally, both positions can be paired, but with different partnersi (3).



Formally, we have the following five categories ($K = 5$): (1) $(i\ bp\ j)$, (2) $(i\ \neg bp\ j)$ & $(i\ bp\ j')$ & $(j\ ss)$, (3) $(i\ \neg bp\ j)$ & $(i\ bp\ j')$ & $(i'\ bp\ j)$, (4) $(i\ \neg bp\ j)$ & $(i\ ss)$ & $(j\ ss)$ and (5) $(i\ \neg bp\ j)$ & $(i\ ss)$ & $(i'\ bp\ j)$ for any pair of bases (i,j) where $i \neq i'$ and $j \neq j'$. Here, $(i\ ss)$ denotes the case that position i is single stranded.

This can be evaluated by the R_K correlation coefficient ($K = 5$) [1]. This correlation coefficient of two assignments represented by two $N \times K$ matrices of data $\underline{\underline{X}}$ and $\underline{\underline{Y}}$ is defined as

$$CC = \frac{COV(\underline{\underline{X}}, \underline{\underline{Y}})}{\sqrt{COV(\underline{\underline{X}}, \underline{\underline{X}})COV(\underline{\underline{Y}}, \underline{\underline{Y}})}}. \tag{1}$$

The covariance between $\underline{\underline{X}}$ and $\underline{\underline{Y}}$ is defined as the expected covariance between the respective k th columns $\underline{\underline{X}}_k$ and $\underline{\underline{Y}}_k$ in the matrices:

$$COV(\underline{\underline{X}}, \underline{\underline{Y}}) = \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} (X_{nk} - \overline{X_k})(Y_{nk} - \overline{Y_k}), \tag{2}$$

where $\overline{X_k} = (1/N) \sum_{n=1}^{N} X_{nk}$ and $\overline{Y_k}$ are the respective means of column k, and X_{nk} are elements of $\underline{\underline{X}}$. Note that Matthews correlation coefficient (MCC) applies to the two categories ($K = 2$) base paired (i bp j) and not base paired (i $\neg bp$ j) for any pair of bases ($N = M(M-1)/2$ where M is length of sequence). Correction for sliding base pairing is not used.

When extending the consideration of unpaired bases, we obtain R_5 correlation coefficients of PETfold: 0.72, Pfold: 0.58, RNAalifold: 0.65. This evaluation is more strict as the two-category Matthews correlation coefficient. Nevertheless, both evaluations show almost the same differences between the three methods.

Detailed Performance Result

SI Table 1 shows the detailed performance listing of PETfold with suggested parameters ($\alpha = 0.2$, $\beta = 1$, $p_{ss}^{\text{threshold}} = 1$ and $p_{bp}^{\text{threshold}} = 0.9$), Pfold and RNAalifold using default parameters. Both Matthews (MCC) and R_5 correlation coefficient (R_5) are listed for the 46 RNA families in the data set. Bold CCs represent the best performance of a family in the 0.01 confidence interval. The alignments are characterized through their number of sequences (#seq), mean pairwise identity (MPI) and number of structural cluster (#cl) calculated by Pcluster. Actually, Pcluster can be improved by using PETfold instead of Pfold. Families in the CMfinder database are indicated by * and high quality alignments documented through the SARSE project are indicated by †. RNA families with the best computational structure prediction (according to MCC) by PETfold are shown at the top, by Pfold in the middle and by RNAalifold at the bottom.

References

[1] J. Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Comput Biol Chem*, 28(5-6):367–74, 2004.

				MCC			R_5		
Family	#seq	MPI	#cl	PETfold	Pfold	RNAalifold	PETfold	Pfold	RNAalifold
†Entero_5_CRE	84	81.50	3	**0.96**	0.83	0.80	**0.92**	0.68	0.56
†HACA_sno_Snake	22	86.85	5	**0.79**	0.25	0.41	**0.59**	0.20	0.12
†HepC_CRE	53	86.93	5	**1**	0.92	0.96	**1**	0.81	0.90
†Hsp90_CRE	3	96.05	1	**0.89**	0.30	0.83	**0.82**	0.10	0.72
†IBV_D-RNA	10	92.91	3	**1**	0.93	0.93	**1**	0.75	0.75
*Lysine	42	54.90	1	**0.94**	0.89	0.91	**0.89**	0.80	0.81
*mir-10	11	57.52	4	**0.89**	0.68	0.76	**0.69**	0.44	0.57
†mir-BART1	2	85.71	1	**0.90**	0.80	0.81	0.60	0.50	**0.62**
†rncO	5	66.19	1	**0.93**	0.53	0.78	**0.81**	0.46	0.58
†SCARNA14	4	80.65	2	**0.75**	0.08	0.70	**0.54**	0.03	0.50
†SNORA18	6	81.14	1	**0.84**	0.67	0.52	**0.63**	0.37	0.33
†SNORD64	2	89.55	2	**0.46**	0.16	0.30	0.14	0.13	**0.17**
†SNORD86	6	75.83	2	**0.50**	0	0	**0.21**	0	0
†TCV_H5	3	94.93	2	**1**	0.73	0.96	**1**	0.58	0.89
†TCV_Pr	4	93.40	2	**1**	0.64	0.96	**1**	0.53	0.62
*Intron_gpII	95	69.09	10	**0.97**	**0.97**	0.79	**0.96**	**0.96**	0.72
*let-7	14	64.63	1	**0.99**	**0.98**	0.90	**0.94**	**0.94**	0.69
*IRE	39	59.48	2	**0.90**	0.84	**0.89**	0.64	0.50	**0.68**
†HCV_SLIV	72	84.31	7	**0.93**	**0.93**	0.90	**0.70**	**0.70**	0.59
*Purine	20	53.25	1	**0.87**	**0.87**	0.85	**0.70**	**0.70**	0.67
*RFN	32	68.79	1	**0.68**	**0.68**	0.53	**0.71**	**0.71**	0.53
*SECIS	62	51.12	1	**0.83**	**0.83**	0.74	**0.73**	**0.73**	0.50
†ykoK	33	61.46	1	**0.86**	**0.87**	0.75	**0.78**	0.75	0.60
*S_box	46	71.31	2	**0.88**	0.86	**0.88**	**0.81**	0.77	**0.80**
†SCARNA15	3	90.03	3	**0.92**	0.61	**0.93**	0.76	0.37	**0.80**
*Histone3	64	77.21	2	**1**	**1**	**1**	**1**	**1**	**1**
*Rhino_CRE	9	79.39	3	**0.71**	**0.71**	**0.72**	0.65	0.56	**0.71**
*Entero_CRE	56	81.72	4	0.87	**0.92**	0.71	0.69	**0.83**	0.52
†HDV_ribozyme	15	89.86	1	0.59	**0.64**	0	**0.32**	0.31	0
†mir-194	4	72.48	2	0.97	**1**	0.76	0.87	**1**	0.57
†RNA-OUT	4	84.29	3	0.74	**0.79**	0.70	**0.50**	**0.51**	0.38
†SNORA38	5	85.68	1	0.70	**0.89**	0.66	0.56	**0.80**	0.50
*Tymo_tRNA-like	28	66.73	2	0.93	**0.95**	0.89	0.78	**0.88**	0.78
†Antizyme_FSE	13	81.83	3	0.89	0.93	**0.96**	0.82	**0.91**	**0.91**
*ctRNA_pGA1	15	72.06	3	0.96	0.96	**0.98**	0.88	0.88	**0.94**
†GcvB	9	54.82	1	0.75	0.73	**0.82**	0.58	0.54	**0.67**
*glmS	10	57.73	1	0.91	0.90	**0.97**	0.81	0.79	**0.91**
*lin-4	9	67.57	5	0.78	0.80	**0.99**	0.72	0.72	**0.93**
†nos_TCE	3	85.94	1	0.92	0.85	**0.98**	0.75	0.53	**0.91**
†Rota_CRE	11	85.91	3	0.64	0.49	**0.76**	0.42	0.25	**0.54**
*s2m	38	78.31	2	0.96	0.78	**1**	0.88	0.77	**1**
†SNORA14	3	92.35	1	0.90	0.78	**0.92**	0.77	0.52	**0.81**
†SNORA40	7	91.37	2	0.84	0.72	**0.95**	0.59	0.50	**0.87**
†SNORA56	2	87.50	4	0.54	0.27	**0.92**	0.50	0.17	**0.85**
†SNORD105	2	77.27	2	0.94	0	**0.97**	0.82	0.15	**0.88**
†snoU83B	4	91.31	1	0.87	0.73	**0.89**	0.66	0.40	**0.71**
AVG	21	77.37	2.4	**0.85**	0.71	0.79	**0.72**	0.58	0.65

Table 1: Detailed performance on data set